Joint Cache Placement and Cooperative Multicast Beamforming in Integrated Satellite-Terrestrial Networks

Dairu Han, Wenhe Liao, Haixia Peng, *Member, IEEE*, Huaqing Wu, *Member, IEEE*, Wen Wu, *Member, IEEE*, and Xuemin Shen, *Fellow, IEEE*

***Abstract*—This paper studies joint cache placement and cooperative multicast beamforming to provide content-centric data services for mobile users in the integrated satellite-terrestrial network (ISTN). Specifically, in the ISTN, users requesting the same content are arranged into a multicast group and served by the cache-enabled base stations (BSs) and low earth orbit (LEO) satellite via cooperative beamforming. To maximize the network utility that takes network throughput and backhaul traffic into consideration, the cache placement, LEO satellite and BS clustering, and multicast beamforming are jointly designed and formulated as a two-timescale optimization problem. However, the original problem is anti-causal since the cache placement strategy and content delivery policy are coupled in different timescales. By utilizing historical information, we propose a two-step scheme to decompose the problem into a short-term content delivery subproblem and a long-term cache placement subproblem. As the former subproblem is nonconvex with mixed-integer variables and coupling constraints, we transform it into an equivalent problem and propose a penalty concave-convex procedure based algorithm to solve it. To address the latter subproblem, a centralized iterative algorithm and a distributed alternating algorithm with low complexity are developed, respectively. Simulation results validate that the proposed schemes can effectively enhance the network throughput and reduce the backhaul traffic compared with benchmark scheme.**

***Index Terms*—Integrated satellite-terrestrial networks, cache placement, content delivery, cooperative multicast beamforming, two-timescale optimization.**

I. Introduction

DRIVEN by various emerging content-centric communications, such as full-motion video streaming, mobile application download, and multimedia content sharing [2], mobile data traffic increases unprecedentedly in recent years. It is predicted that the global mobile data traffic will reach 226 EB per month in 2026, growing by a factor of around 4.5 compared with that in 2020 [3]. The unparalleled mobile data growth brings huge challenges for conventional cellular networks in providing high throughput and multi-accessibility services with limited backhaul capacity. To cope with this challenge, edge caching is a potential approach to relieve the backhaul burden in terrestrial networks [4]. By proactively caching frequently requested contents at edge nodes, popular contents can be directly delivered to users without fetching contents from content servers via backhaul links, which efficiently avoids massive repetitive content fetching and significantly alleviates backhaul pressure [5]. However, the traditional cache-enabled terrestrial network cannot cope with the expected skyrocketing data increase, owing to the constrained spectrum resource and upper bound of energy efficiency in terrestrial networks [6].

With the merits of broadcast transmission and tremendous coverage, satellite networks are considered as the complement to terrestrial networks, especially in cached content delivery [7]. Recently, the rapid development of satellite launch and miniaturization technologies facilitates the economy-friendly commercial deployment of low earth orbit (LEO) satellites [8]. By combining edge caching and satellite communication, a potentially unlimited number of users can be served across a single beam transmission [9]. Terrestrial networks can provide users with high-speed data transmission services at low cost, and satellite networks can serve users with broad service coverage. The integration of both networks, i.e., the cache-enabled integrated satellite-terrestrial network (ISTN), is envisioned to be a promising architecture in the next generation wireless networks to facilitate ubiquitous and flexible data transmission services [10]. In the cache-enabled ISTN, content caching generally consists of two stages, i.e., cache placement and content delivery. In the *cache placement* stage, since the cached contents at edge nodes are updated in a long-term process [11], the cache placement strategy should be judiciously devised to improve content hit ratio. In the *content delivery* stage, it is essential to dynamically design the content delivery policy to transmit the required contents to mobile users in a short-term process.

Manuscript received July 31, 2021; revised November 29, 2021; accepted December 15, 2021. Date of publication December 28, 2021; date of current version March 15, 2022. This paper was presented in part at the IEEE Conference on Communications 2021 [1]. This work was supported in part by China Scholarship Council and in part by the Engineering Research Council (NSERC) of Canada. The review of this article was coordinated by Dr. Haijun Zhang. *(Corresponding author: Wenhe Liao.)*

Dairu Han and Wenhe Liao are with the Department of Aeronautical and Astronautical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail: handairu@njust.edu.cn; cnwho@njust.edu.cn).

Haixia Peng is with the Department of Computer Engineering and Computer Science, California State University Long Beach, Long Beach, CA 90840 USA (e-mail: h27peng@uwaterloo.ca).

Huaqing Wu and Xuemin Shen are with the Electrical and Computer Engineering, University of Waterloo, Waterloo N2L 3G1, Canada (e-mail: h272wu@uwaterloo.ca; sshen@uwaterloo.ca).

Wen Wu is with the Frontier Research Center, Peng Cheng Laboratory, Shenzhen 518066, Guangdong, China (e-mail: w77wu@uwaterloo.ca).

Digital Object Identifier 10.1109/TVT.2021.3138898

Considering that popular contents may be requested by multiple mobile users simultaneously, cooperative multicast beamforming is an effective delivery approach in the cache-enabled ISTN [12]. With cooperative multicast beamforming, the requests of multiple mobile users can be met using the same resource block, therefore efficiently reduces the repetitive transmissions and improves spectral efficiency [13]. Meanwhile, users' received signal-to-interference-plus-noise ratio (SINR) can be improved due to the cooperative beamforming among the satellites and terrestrial base stations (BSs) [14], [15]. Nevertheless, performing full cooperative multicast beamforming requires content sharing among all BSs and satellites. Via satellite and BS clustering, the contents only need to be delivered by a cluster of serving satellites and BSs, which further alleviates backhaul link pressure [16].

Despite the advantages of combining edge caching and multicast beamforming in the ISTN, many challenges persist in jointly designing the long-term cache placement strategy and short-term content delivery policy (i.e., LEO satellite and BS clustering and multicast beamforming) since the two stages are tightly coupled. For one thing, the update of cache placement should cater for content delivery to enhance the utilization of caching resources. It is non-trivial to rationally devise the long-term cache placement strategy to satisfy future requests without prior information on the content popularity distribution and channel conditions. For another, the short-term content delivery policy, especially the cooperative pattern of satellites and BSs, is not only affected by channel conditions but also determined by the cache status at BSs and satellites. It is necessary to develop an efficient content delivery policy that is both channel-aware and cache-aware.

To maximize the network utility which takes both network throughput and total backhaul traffic into consideration, we jointly design the cache placement strategy and content delivery policy in the cache-enabled ISTN while addressing the above challenges. The main contributions of this paper are summarized as follows.

- We introduce a novel cache-enabled ISTN to provide content-centric data services for mobile users, where an LEO satellite and BSs cooperatively serve mobile users through cooperative multicast beamforming.
- We formulate a two-timescale optimization problem to maximize the network utility considering the long-term average network throughput and backhaul traffic by jointly optimizing the cache placement, BS and satellite clustering, and multicast beamforming. Since this problem is anti-causal with two timescales, by utilizing the historical users' request information and stochastic channel distribution information, we propose a two-step scheme to decompose the original optimization problem into a short-term content delivery subproblem and a long-term cache placement subproblem.
- We propose a penalty concave-convex procedure (P-CCCP) based algorithm to deal with the short-term content delivery subproblem, which is a mixed integer nonlinear programming (MINLP) problem due to the coupling of the continuous and discrete variables.
- We propose a centralized iterative algorithm and a distributed alternating algorithm based on the block coordinate descent (BCD) method to solve the long-term cache placement subproblem. We analyze the computational complexity of the two proposed algorithms and compare their respective advantages.

The remainder of this paper is organized as follows. We review the related works in Section II and describe the system model in Section III. The problem formulation and decomposition are presented in Section IV. In Section V, the P-CCCP based algorithm is proposed to solve the short-term content delivery subproblem. We elaborate the proposed centralized and distributed algorithms for the long-term cache placement subproblem in Section VI. The performance evaluation is conducted in Section VII. Finally, we conclude this paper in Section VIII.

II. Related Work

The cache placement and content delivery have attracted great attention recently. For cache placement design, in [17], a cooperative hierarchical caching framework was proposed to minimize the average delay-cost of content delivery in a cloud radio access network, in which a new cache-enabled central processor (CP) was introduced. By utilizing separate channels for content dissemination, Sung *et al.* [18] presented an efficient cache placement strategy to reduce the service delay and improve packet delivery ratio in a two-tier network. To rationally devise the content delivery policy given a cache placement strategy, a content-centric transmission design incorporated with multicasting and BS clustering was investigated in [12] to minimize the transmission cost. Assuming that new users and contents randomly arrive into the network, Qin *et al.* [19] studied how to maximize the content delivery rate in wireless networks. To identify the relation between cache placement and content delivery, the cache placement, multicast beamforming, and BS clustering were jointly optimized to reduce the system energy consumption and improve cache efficiency in [11] and [20].

With the development of satellite on-board caching capacity and software defined networking (SDN) and network function virtualization (NFV) techniques, remarkable papers have considered the ISTN and cache-enabled LEO satellites to provide ubiquitous and resilient content delivery services. In [10], LEO satellites were integrated with the non-orthogonal multiple access based terrestrial networks to cooperatively serve different users aiming at increasing the system throughput, where satellites only provide services for users far away from BSs. To fully explore the cooperation between the satellite and terrestrial networks, CP coordinates satellites and BSs to cooperatively serve all the users through joint multicast beamforming in [14]. Taking the constraint of backhaul link capacity into account, Zhang *et al.* [16] further proposed a joint beamforming and resource allocation scheme to satisfy users' QoE in the ISTN. To alleviate the transmission delay and leverage satellite's immense coverage, in-network caching gives a new impulse to the ISTN. Yang *et al.* [21] proposed a time-evolving covering set instructed caching method to degrade the overheads and access delay of users in the ISTN. Considering the dynamic satellite link and

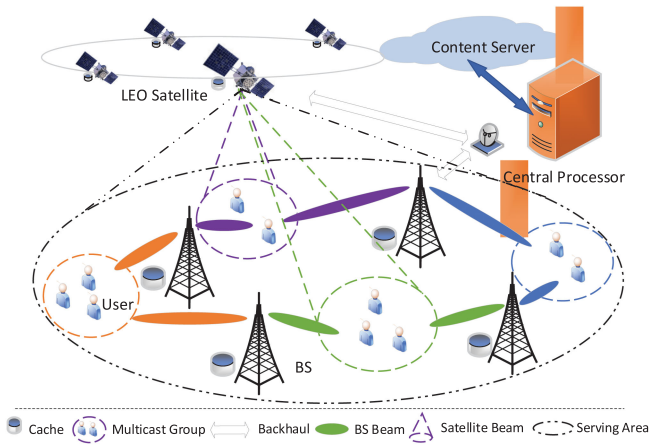


Fig. 1. The system model of th'e considered cache-enabled ISTN.

topology, a QoE-aware content distribution scheme for ISTN was presented in [22], by which a stable user experience can be achieved.

Different from the existing works, we jointly design the cache placement strategy and content delivery policy in the ISTN. In [1], we investigated joint cache placement and content delivery to minimize the energy consumption in the ISTN. In this paper, with the objective of optimizing two intertwined network utilities (i.e., network throughput and total backhaul traffic), we formulate a two-timescale optimization problem. To solve the problem, new algorithms are proposed for the decoupled two subproblems. For the small timescale subproblem, we propose a P-CCCP based algorithm, which achieves a lower computational complexity and a better performance than the one in [1]. For the large timescale subproblem, we propose a centralized algorithm and a distributed algorithm. In addition, we theoretically analyze their computational complexity and advantages.

III. System Model

In this section, we first describe the cache-enabled ISTN model. Then, the detailed cache and user request models and transmission model are introduced.

A. Network Model

As depicted in Fig. 1, we consider the downlink of a cache-enabled ISTN consisting of several multi-antenna LEO satellites, B multi-antenna BSs, N single-antenna mobile users, and one CP. We consider each satellite has non-overlapping coverage with other satellites and periodically provides services to a target area for a certain duration to guarantee seamless coverage. Without loss of generality, we focus on the scenario where one satellite and B BSs cooperatively provide the content delivery services to randomly distributed users in the target area.[1] Each BS and satellite have M_b and M_s antennas, respectively. All the BSs and the satellite have a limited cache storage capacity and are connected to the CP via high-speed backhaul links.

[1]Note that the handover process of the satellites can be managed by the CP, which is beyond the scope of this work [23].

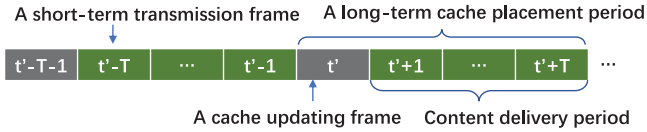


Fig. 2. The two-timescale structure of the considered system.

For simplicity, BSs and the satellite are hereinafter referred to as access nodes (ANs). Denote the index of ANs by $b \in \mathcal{B} = \{0, 1, 2, \ldots, B\}$, where $b = 0$ represents the satellite, and $1 \leq b \leq B$ indicates the BSs. Let $\mathcal{N} = \{1, 2, \ldots, N\}$ denote the set of users. Under the schedule of the CP,[2] all users can be cooperatively served by the associated cluster of ANs by joint multicasting while reusing the entire spectrum. The CP can access the content server for a library of $\mathcal{F} = \{1, 2, \ldots, F\}$ in which F contents are potentially requested by users.

In the considered system, the content delivery policy is determined by the CP in each transmission frame t. Specifically, the wireless channel is considered to be static during each transmission frame and varies over different transmission frames. Note that in practice, the content popularity distribution generally varies much slower than the wireless channel condition [25]. Moreover, the cache placement may not be frequently updated in order to reduce the complexity and potential communication overhead. Therefore, the cache placement is normally updated in a long-term period which consists of many short-term transmission frames. To elaborate the cache placement and content delivery process, we consider a two-timescale structure, including long-term cache placement periods and short-term transmission frames, as shown in Fig. 2. Specifically, in each long-term cache placement period, the cache placement is updated at the beginning of frame t' and remains unchanged for a content delivery period $\mathcal{T} = \{t' + 1, t' + 2, \ldots, t' + T\}$, which includes T short-term transmission frames. In each short-term transmission frame, the CP jointly designs beamformers and serving clusters for ANs according to users' requests, instantaneous channel state information (CSI), and the current caching status. The instantaneous CSI can be acquired in the CP via either model-based or data-driven prediction methods [26]. If the users' requested contents have been cached in the associated AN cluster, the ANs can deliver the requested contents to the users directly. Otherwise, the ANs need to fetch the missing data from the content server and then deliver the contents.

B. Cache and User Request Models

The content popularity distribution can be characterized by the classical Zipf distribution which has been shown to fit well with measured web and video requests [27]. Specifically, the content popularity of the f-th content is given by

$$p_f = \frac{f^{-\delta}}{\sum_{i=1}^{F} i^{-\delta}}, \quad \forall f \in \mathcal{F}, \tag{1}$$

[2]With the development of the SDN and cloud radio access network technology, it is possible to control BSs and the satellite in a centralized manner by the CP for joint multicasting and interference management [14], [24].

where δ is the popularity skewness characterizing the user preference of the contents, and f denotes the content popularity rank order. A larger value of popularity skewness means that content requests are more concentrated. Different from the previous works that consider homogenous content popularity distribution [28], [29], we consider heterogeneous content popularity distribution for users in this paper, i.e., different users have different content preferences.

During cache updating frame t', the cached contents in each AN should be periodically updated to enhance the utilization of caching resources. We denote $\mathbf{A} = [a_{f,b}]^{F\times(B+1)}$ as the cache placement matrix for all ANs, where $a_{f,b} \in [0,1]$ represents the fraction of content f cached at the b-th AN.[3] In addition, due to the limited cache storage capacity of ANs, we have $\sum_{f=1}^{F} a_{f,b} m_f \leq x_b \sum_{f=1}^{F} m_f$, where x_b is the fractional cache capacity of AN b, i.e., the maximum fraction of the F contents that can be cached by AN b, and m_f is the data size of content f.

C. Transmission Model

At the beginning of each short-term transmission frame, each user requests one desired content based on individual content preference. The requests from all users in frame t are denoted by $\mathcal{K}_t = \{k_{1,t}, \ldots k_{n,t}, \ldots, k_{N,t}\}$, where $k_{n,t} = f, \forall f \in \mathcal{F}$ means that user n requests content f in frame t. According to the content requests, users requesting the same content are scheduled into the same multicast group which are served by a selected cluster of ANs via joint beamforming. The number of formed groups and the set of requested contents in frame t are denoted by F_t and $\mathcal{F}_t = \{1, 2, \ldots, F_t\}$, respectively. Specifically, users requesting the f-th content during frame t are grouped together, denoted by $\mathcal{G}_{f,t}, \forall f \in \mathcal{F}_t$.

Let $\mathbf{w}_{f,b,t} \in \mathbb{C}^{M_b\times 1}$ denote the downlink beamforming vector from BS b to multicast group $\mathcal{G}_{f,t}$. Then, the signal transmitted by BS b at frame t is $\mathbf{S}_{b,t} = \sum_{f\in\mathcal{F}_t} \mathbf{w}_{f,b,t} s_{f,t}, \forall b \in \mathcal{B}\backslash\{0\}$, where $s_{f,t} \in \mathbb{C}$ is data symbol of the content requested by group $\mathcal{G}_{f,t}$ with $\mathbb{E}[|s_{f,t}|^2] = 1$. Similarly, the signal transmitted by the satellite at frame t is expressed as $\mathbf{S}_{s,t} = \sum_{f\in\mathcal{F}_t} \mathbf{v}_{f,t} s_{f,t}$, where $\mathbf{v}_{f,t} \in \mathbb{C}^{M_s\times 1}$ denotes the downlink beamforming vector from the satellite for multicast group $\mathcal{G}_{f,t}$. Subsequently, the signal received at user n in multicast group $\mathcal{G}_{f,t}$ can be expressed as

$$\begin{aligned} y_{n,f,t} = &\sum_{b=1}^{B} \mathbf{h}_{n,b,t}^H \mathbf{w}_{f,b,t} s_{f,t} + \mathbf{g}_{n,t}^H \mathbf{v}_{f,t} s_{f,t} \\ &+ \sum_{b=1}^{B} \sum_{f'\in\mathcal{F}_t\backslash\{f\}} \mathbf{h}_{n,b,t}^H \mathbf{w}_{f',b,t} s_{f',t} \\ &+ \sum_{f'\in\mathcal{F}_t\backslash\{f\}} \mathbf{g}_{n,t}^H \mathbf{v}_{f',t} s_{f',t} + n_{n,t}, \end{aligned} \tag{2}$$

where $\mathbf{h}_{n,b,t} \in \mathbb{C}^{M_b\times 1}$ is the channel vector from BS b to user n at frame t, $\mathbf{g}_{n,t} \in \mathbb{C}^{M_s\times 1}$ denotes the channel vector from the satellite to user n at frame t, and $n_{n,t} \sim \mathcal{CN}(0, \sigma_k^2)$ represents the additive white Gaussian noise. To make the scenario more realistic, the terrestrial channel is modeled by the Rayleigh channel with shadowing effect. Both the large-scale fading and the shadowed-Rician fading are considered to model channels between the satellite and users [30]. We denote $\mathbf{h}_{n,t} = [\mathbf{h}_{n,1,t}^H, \mathbf{h}_{n,2,t}^H, \ldots, \mathbf{h}_{n,B,t}^H]^H \in \mathbb{C}^{BM_b\times 1}$ as the aggregated channel vectors from all BSs to user n. Let $\mathbf{w}_{f,t} = [\mathbf{w}_{f,1,t}^H, \mathbf{w}_{f,2,t}^H, \ldots, \mathbf{w}_{f,B,t}^H]^H \in \mathbb{C}^{BM_b\times 1}$ denote the aggregated beamforming vectors from all BSs to group $\mathcal{G}_{f,t}$. Then, the received SINR of user n in group $\mathcal{G}_{f,t}$ is given by

$$\gamma_{n,f,t} = \frac{|\mathbf{h}_{n,t}^H \mathbf{w}_{f,t} + \mathbf{g}_{n,t}^H \mathbf{v}_{f,t}|^2}{\sum_{f'\in\mathcal{F}_t\backslash\{f\}} |\mathbf{h}_{n,t}^H \mathbf{w}_{f,'t} + \mathbf{g}_{n,t}^H \mathbf{v}_{f,'t}|^2 + \sigma_N^2}. \tag{3}$$

In a multicast scenario, the transmission rate of a multicasting group is determined by the user with the lowest SINR [31]. Hence, the transmit rate of multicasting group $\mathcal{G}_{f,t}$ is given by

$$R_{f,t} = B_0 \log_2 \left(1 + \min\{\gamma_{n,f,t}\}\right), \forall n \in \mathcal{G}_{f,t}, f \in \mathcal{F}_t, \tag{4}$$

where B_0 is the channel bandwidth. The sum network throughput can be expressed as $\sum_{f\in\mathcal{F}_t} R_{f,t}$. In addition, we have the following transmit power constraints for each BS and the satellite, i.e.,

$$\begin{aligned} &\sum_{f\in\mathcal{F}_t} ||\mathbf{w}_{f,b,t}||_2^2 \leq P_B^{max}, \quad \forall b \in \mathcal{B}\backslash\{0\}, \\ &\sum_{f\in\mathcal{F}_t} ||\mathbf{v}_{f,t}||_2^2 \leq P_S^{max}, \end{aligned} \tag{5}$$

where P_B^{max} and P_S^{max} are the maximum transmit powers of BSs and the satellite, respectively.

To enhance the network throughput, a cluster of ANs is assigned to each multicast group. We define the AN clustering matrix at frame t as

$$\mathbf{C}_t = [c_{f,b,t}] \in \{0,1\}^{F_t\times(B+1)}, \quad \forall b \in \mathcal{B}, f \in \mathcal{F}_t. \tag{6}$$

Here, $c_{f,b,t} = 1$ indicates that AN b is scheduled to serve multicast group $\mathcal{G}_{f,t}$ and $c_{f,b,t} = 0$ otherwise. When $c_{f,b,t} = 0$, the corresponding beamforming vector from AN b to multicast group $\mathcal{G}_{f,t}$ is $\mathbf{0}$. In this case, the following constraints should be satisfied:

$$\begin{aligned} &(1 - c_{f,b,t})\mathbf{w}_{f,b,t} = \mathbf{0}, \quad \forall b \in \mathcal{B}\backslash\{0\}, f \in \mathcal{F}_t, \\ &(1 - c_{f,b,t})\mathbf{v}_{f,t} = \mathbf{0}, \quad b = 0, \ \forall f \in \mathcal{F}_t. \end{aligned} \tag{7}$$

In each transmission frame, if the contents requested by multicast group $\mathcal{G}_{f,t}$ is not available in the associated ANs, the ANs should first fetch the missing content segment from the content server and then deliver the requested contents. In this case, the backhaul traffic consumed by AN b in transmission frame t is expressed by

$$R_{b,t}^{BH} = \sum_{f\in\mathcal{F}_t} c_{f,b,t}(1 - a_{f,b}) m_f. \tag{8}$$

[3] In light of the limited computing capabilities of the LEO satellites, we consider uncoded caching in the ISTN rather than coded caching.

IV. Problem Formulation and Decomposition

In this section, we first formulate a two-timescale optimization problem to maximize the network utility through jointly designing cache placement, multicast beamforming, and AN clustering. Then, a two-step scheme is proposed to decompose the original problem into two tractable subproblems.

A. Problem Formulation

In our framework, both cache placement and content delivery impact network performance. On the one hand, if the BSs and the satellite that cache the requested contents and have poor channel conditions are clustered to jointly serve a specific user, the backhaul traffic can be mitigated while the network throughput degrades. On the other hand, if the selected BSs and satellite have good channel conditions and the requested contents are miss cached, the network throughput can be improved while the backhaul traffic increases. Therefore, the short-term transmission policy (including beamformer design and AN clustering) and the long-term cache placement strategy are tightly coupled. Moreover, the network utilities of network throughput and backhaul traffic are intertwined. Therefore, it is essential to jointly design cache placement strategy and content delivery policy to strike a balance between network throughput improvement and backhaul traffic reduction.

In this paper, we aim to maximize the long-term average network utility that takes both network throughput and total backhaul traffic into consideration. An optimization problem $\mathcal{P}$ is formulated, in which the long-term variable (cache placement matrix $\mathbf{A}$) and the short-term variables (multicast beamformers $\mathbf{w}_t = [\mathbf{w}_{1,t}^H, \mathbf{w}_{2,t}^H, \ldots, \mathbf{w}_{F_t,t}^H]^H \in \mathbb{C}^{F_t B M_b \times 1}$ and $\mathbf{v}_t = [\mathbf{v}_{1,t}^H, \mathbf{v}_{2,t}^H, \ldots, \mathbf{v}_{F_t,t}^H]^H \in \mathbb{C}^{F_t M_s \times 1}$, and AN clustering matrix $\mathbf{C}_t$) are jointly optimized. Problem $\mathcal{P}$ is given by

$$\mathcal{P}: \max_{\{\mathbf{C}_t, \mathbf{w}_t, \mathbf{v}_t\}, \mathbf{A}} \frac{1}{T} \sum_{t \in \mathcal{T}} \left(\sum_{f \in \mathcal{F}_t} R_{f,t} - \beta \sum_{b=0}^{B} R_{b,t}^{BH} \right) \tag{9}$$

$$\text{s.t. } a_{f,b} \in [0,1], \quad \forall b \in \mathcal{B}, f \in \mathcal{F}, \tag{9a}$$

$$\sum_{f=1}^{F} a_{f,b} m_f \le x_b, \quad \forall b \in \mathcal{B}, \tag{9b}$$

$$(5)-(7), \quad \forall t \in \mathcal{T}, \tag{9c}$$

where β is the trade-off parameter to reflect the preference between network throughput maximization and backhaul traffic minimization. Specifically, a larger value of β indicates that reducing backhaul traffic is more important than increasing network throughput in the considered system. As the backhaul traffic consumed for cache updating is negligible compared with that for content transmission in content delivery period $\mathcal{T}$, we do not consider the consumed backhaul traffic at cache updating frame t' in the objective function.

There exist two main challenges in solving problem $\mathcal{P}$. First, as elaborated in Subsection III-A, cache placement is designed at frame t' and should be adaptive to the users' requests $\{\mathcal{K}_t, \forall t \in \mathcal{T}\}$ and the channel vectors $\{\mathbf{h}_{n,t}\}$ and $\{\mathbf{g}_{n,t}\}$ in the subsequent T transmission frames. However, accurately knowing the users' requests and channel vectors in following transmission frames is generally unpractical. Moreover, multicast beamformers and AN clustering matrix are designed based on the cache placement matrix, $\mathbf{A}$. Therefore, problem $\mathcal{P}$ is anti-causal when matrix $\mathbf{A}$ is updated at frame t'. Second, the short-term variables $\{\mathbf{w}_t\}$, $\{\mathbf{v}_t\}$, and $\{\mathbf{C}_t\}$ are mutually coupled with long-term variable $\mathbf{A}$ in the objective function which further complicates the optimization problem. Thus, it is painstaking to solve problem $\mathcal{P}$.

B. Problem Decomposition

In practice, channel statistics are slowly varying and content popularity is relatively unchanged in two consecutive long-term cache placement periods. Consequently, the cache placement strategy for future users' requests can be predicted by utilizing historical users' request information and CSI [20]. Therefore, we propose a two-step cache placement and content delivery scheme to decompose problem $\mathcal{P}$ into two subproblems.

In the cache placement step, the historical users' request information and CSI in the last content delivery period, i.e., $\mathcal{T}' = \{t'-T, t'-T+1, \ldots, t'-1\}$, are utilized to design the optimal long-term cache placement matrix, $\mathbf{A}^*$, at current cache updating frame t'. Then, in the content delivery step, with optimal cache placement matrix $\mathbf{A}^*$, we maximize the objective function in problem $\mathcal{P}$ by optimizing the short-term variables $\mathbf{w}_t$, $\mathbf{v}_t$, and $\mathbf{C}_t$ in each transmission frame t during current content delivery period $\mathcal{T} = \{t'+1, t'+2, \ldots, t'+T\}$. These two subproblems are given as follows.

1) Long-Term Cache Placement Subproblem: With the information of the historical users' requests and CSI, the long-term cache placement matrix $\mathbf{A}$ at cache updating frame t' can be optimized by solving the following problem:

$$\mathcal{P}_L: \max_{\{\mathbf{C}_t, \mathbf{w}_t, \mathbf{v}_t\}, \mathbf{A}} \frac{1}{T} \sum_{t \in \mathcal{T}'} \left(\sum_{f \in \mathcal{F}_t} R_{f,t} - \beta \sum_{b=0}^{B} R_{b,t}^{BH} \right)$$

$$\text{s.t. } (5)-(7), \quad \forall t \in \mathcal{T}',$$

$$(9a)-(9b).$$

Remark 1: Note that the main difference between problems $\mathcal{P}$ and $\mathcal{P}_L$ is that the historical users' request information and CSI[4] during content delivery period $\mathcal{T}'$ are used in problem $\mathcal{P}_L$. In addition, short-term variables $\{\mathbf{w}_t, \mathbf{v}_t, \mathbf{C}_t, \forall t \in \mathcal{T}'\}$ also need to be optimized to obtain the optimal cache placement matrix, $\mathbf{A}^*$, where $\mathbf{A}^*$ is the only output of problem $\mathcal{P}_L$.

2) Short-Term Content Delivery Subproblem: Since the content delivery policy in each short-term transmission frame is independent, the short-term variables ($\mathbf{w}_t$, $\mathbf{v}_t$, and $\mathbf{C}_t$, for $t = t'+1, t'+2, \ldots, t'+T$) are optimized by solving the following problem:

$$\mathcal{P}_S: \max_{\mathbf{C}_t, \mathbf{w}_t, \mathbf{v}_t} \sum_{f \in \mathcal{F}_t} R_{f,t} - \beta \sum_{b=0}^{B} R_{b,t}^{BH}$$

$$\text{s.t. } (5)-(7).$$

[4]The historical users' request information and CSI are available in the CP at cache updating frame t'.

Here, $\mathcal{P}_L$ and $\mathcal{P}_S$ are both nonconvex and MINLP problems, due to the coupling between the continuous beamforming vectors ($\mathbf{w}_t$ and $\mathbf{v}_t$) and the discrete AN clustering matrix, $\mathbf{C}_t$, in nonconvex constraint (7). Furthermore, for problem $\mathcal{P}_L$, the objective function is nonconvex, and the binary cache placement matrix $\mathbf{A}$ and AN clustering matrix $\mathbf{C}_t$ are coupled with each other in the objective function. Hence, attaining the global optimal solutions for problems $\mathcal{P}_L$ and $\mathcal{P}_S$ is NP-hard and non-tractable [32]. In this paper, our goal is to effectively solve problems $\mathcal{P}_L$ and $\mathcal{P}_S$ in polynomial time. Since the solution to problem $\mathcal{P}_L$ is based on the solution to problem $\mathcal{P}_S$, in the following, we first tackle problem $\mathcal{P}_S$ in Section V and then solve problem $\mathcal{P}_L$ in Section VI.

V. Proposed Algorithm for Short-Term Content Delivery Subproblem

In this section, we propose an efficient algorithm to solve the short-term content delivery subproblem $\mathcal{P}_S$. To make problem $\mathcal{P}_S$ tractable, we first reformulate it by introducing some auxiliary variables and then present a P-CCCP based algorithm to iteratively solve the reformulated problem.

A. Problem Reformulation

We first define two auxiliary variables, $\mathbf{j}_{f,t} \in \mathbb{C}^{(M_s+BM_b)\times 1}$ and $\mathbf{d}_{n,t} \in \mathbb{C}^{(M_s+BM_b)\times 1}$, i.e.,

$$\begin{aligned}\mathbf{j}_{f,t} &= [\mathbf{v}_{f,t}^H, \mathbf{w}_{f,1,t}^H, \mathbf{w}_{f,2,t}^H, \ldots, \mathbf{w}_{f,B,t}^H]^H,\ \forall f \in \mathcal{F}_t,\\ \mathbf{d}_{n,t} &= [\mathbf{g}_{n,t}^H, \mathbf{h}_{n,1,t}^H, \mathbf{h}_{n,2,t}^H, \ldots, \mathbf{h}_{n,B,t}^H]^H,\ \forall n.\end{aligned} \tag{10}$$

In order to represent $\mathbf{v}_{f,t}$ and $\mathbf{w}_{f,b,t}$ with $\mathbf{j}_{f,t}$, we further introduce a set of binary diagonal selection matrices $\{\mathbf{S}_b\}_{b=0}^B$, where $\mathbf{S}_b \in \{0,1\}^{(M_s+BM_b)\times(M_s+BM_b)}$ is defined as

$$\mathbf{S}_b = \begin{cases} \text{diag}([\mathbf{1}_{M_s}^H, \mathbf{0}_{BM_b}^H]), & \text{if } b = 0,\\ \text{diag}([\mathbf{0}_{M_s}^H, \mathbf{0}_{(b-1)M_b}^H, \mathbf{1}_{M_b}^H, \mathbf{0}_{(B-b)M_b}^H]), & \text{otherwise.}\end{cases}$$

Here, $\mathbf{1}_{M_s}^H$ denotes a M_s-dimension all-ones vector, and $\mathbf{0}_{BM_b}^H$ denotes a BM_b-dimension all-zeros vector. Accordingly, constraints (5) and (7) can be equivalently rewritten as

$$\sum_{f\in\mathcal{F}_t} ||\mathbf{j}_{f,t}\mathbf{S}_b||_2^2 \le P_b^{max}, \quad \forall b \in \mathcal{B}, \tag{11}$$

and

$$(1 - c_{f,b,t})\mathbf{j}_{f,t}\mathbf{S}_b = \mathbf{0}, \quad \forall b \in \mathcal{B},\ f \in \mathcal{F}_t, \tag{12}$$

respectively. Here, $P_b^{max} = P_S^{max}$ when $b = 0$, otherwise $P_b^{max} = P_B^{max}$.

Meanwhile, $\mathcal{P}_S$ is a max-min optimization problem and is hard to be solved directly. Thus, we replace the minimum operators $\min\{\gamma_{n,f,t}\}$ by introducing auxiliary variables $\{\gamma_{f,t}\}$, and then problem $\mathcal{P}_S$ can be equivalently reformulated as

$$\mathcal{P}_S1: \max_{\mathbf{C}_t,\mathbf{j}_t,\{\gamma_{f,t}\}} \sum_{f\in\mathcal{F}_t} R'_{f,t} - \beta\sum_{b=0}^{B} R_{b,t}^{BH} \tag{13}$$

$$\text{s.t. } \frac{|\mathbf{d}_{n,t}^H\,\mathbf{j}_{f,t}|^2}{\sum_{f'\in\mathcal{F}_t\setminus\{f\}} |\mathbf{d}_{n,t}^H\,\mathbf{j}_{f,'t}|^2 + \sigma_N^2} \ge \gamma_{f,t},$$

$$\forall f \in \mathcal{F}_t,\ n \in \mathcal{G}_{f,t}, \tag{13a}$$

$$(6), (11), \text{ and } (12), \tag{13b}$$

where $\mathbf{j}_t = [\mathbf{j}_{1,t}^H, \mathbf{j}_{2,t}^H, \ldots, \mathbf{j}_{F_t,t}^H]^H \in \mathbb{C}^{F_t(M_s+BM_b)\times 1}$ and $R'_{f,t} = B_0\log_2(1+\gamma_{f,t})$. Constraint (13a) represents that the SINR of all users in group $\mathcal{G}_{f,t}$ should be larger than the minimum SINR requirement, $\gamma_{f,t}$.

Note that (6) is a binary constraint and (12) is a nonconvex nonlinear equality constraint which both are difficult to tackle. Thus, constraint (6) can be equivalently rewritten as two continuous constraints, i.e.,

$$c_{f,b,t} - c_{f,b,t}^2 \le 0, \forall f \in \mathcal{F}_t, b \in \mathcal{B}, \tag{14}$$

$$c_{f,b,t}^2 - c_{f,b,t} \le 0, \forall f \in \mathcal{F}_t, b \in \mathcal{B}. \tag{15}$$

Constraint (14) is nonconvex and its left-hand side is a difference of two convex (DC) functions. Thus, the CCCP method can be implemented to efficiently tackle the nonconvexity of constraint (14). Meanwhile, considering constraint (11), nonlinear equality constraint (12) can be equivalently represented as the following convex constraint:

$$||\mathbf{j}_{f,t}\mathbf{S}_b||_2 \le c_{f,b,t}\sqrt{P_b^{max}}, \forall f \in \mathcal{F}_t, b \in \mathcal{B}. \tag{16}$$

It is clear that constraints (12) and (16) are equivalent, where associated beamformers is $\mathbf{0}$ for $c_{f,b,t} = 0$.

To deal with nonconvex constraint (13a), we further introduce slack variables $x_{n,t}$, and then problem $\mathcal{P}_S1$ can be reformulated as

$$\mathcal{P}_S2: \max_{\substack{\mathbf{C}_t,\mathbf{j}_t,\\ \{\gamma_{f,t}\},\{x_{n,t}\}}} \sum_{f\in\mathcal{F}_t} R'_{f,t} - \beta\sum_{b=0}^{B} R_{b,t}^{BH} \tag{17}$$

$$\text{s.t. } \gamma_{f,t} - \frac{|\mathbf{d}_{n,t}^H\,\mathbf{j}_{f,t}|^2}{x_{n,t}} \le 0, \forall f \in \mathcal{F}_t, n \in \mathcal{G}_{f,t}, \tag{17a}$$

$$\sum\nolimits_{f'\in\mathcal{F}_t\setminus\{f\}} |\mathbf{d}_{n,t}^H\,\mathbf{j}_{f,'t}|^2 + \sigma_N^2 \le x_{n,t},$$

$$\forall f \in \mathcal{F}_t,\ n \in \mathcal{G}_{f,t}, \tag{17b}$$

$$(11), (14)-(16), \tag{17c}$$

where constraint (13a) is replaced by constraints (17a) and (17b). The equivalence of problems $\mathcal{P}_S1$ and $\mathcal{P}_S2$ can be easily verified when constraint (17b) holds equality at the optimum of problem $\mathcal{P}_S2$. It can be seen from problem $\mathcal{P}_S2$ that constraints (14) and (17a) are still nonconvex and hard to be converted into equivalent convex ones. In the following, we propose a P-CCCP based algorithm to efficiently address the nonconvexity issue and solve problem $\mathcal{P}_S2$.

B. Proposed P-CCCP Based Algorithm

Recall that constraint (14) is expressed as a DC function. Therefore, we resort to the CCCP method to iteratively deal with problem $\mathcal{P}_S2$. It is worth mentioning that finding a feasible initial point is required for employing the CCCP method. However, it

is a non-trivial initialization for the nonconvex problem $\mathcal{P}_S2$. In addition, it is also strenuous to escape from the initial point because of the tightly coupled constraints (14) and (15). To bypass this difficulty of initialization, we first introduce auxiliary variables $\{q_{f,b,t}\}$ to relax constraint (14) by satisfying

$$c_{f,b,t} - c_{f,b,t}^2 \leq q_{f,b,t}, \forall f \in \mathcal{F}_t,\ b \in \mathcal{B}, \tag{18}$$

$$q_{f,b,t} \geq 0, \forall f \in \mathcal{F}_t,\ b \in \mathcal{B}. \tag{19}$$

Accordingly, we then penalize constraint (18) into the objective function to minimize the violation. Thus, problem $\mathcal{P}_S2$ can be transformed into

$$\mathcal{P}_S3: \max_{\substack{\mathbf{C}_t,\mathbf{j}_t,\{q_{f,b,t}\}\\ \{\gamma_{f,t}\},\{x_{n,t}\}}} \sum_{f\in\mathcal{F}_t} R'_{f,t} - \beta\sum_{b=0}^{B} R_{b,t}^{BH} - \lambda\sum_{b=0}^{B}\sum_{f\in\mathcal{F}_t} q_{f,b,t}$$

$$\text{s.t. } (11), (15), (16), (17\text{a}), (17\text{b}), (18), \text{ and } (19),$$

where λ is a nonnegative penalty parameter. Each variable $c_{f,b,t}$ acquires a binary value for $q_{f,b,t} = 0$, and increasing λ will force $q_{f,b,t}$ to 0, in which the equality of problems $\mathcal{P}_S2$ and $\mathcal{P}_S3$ is achieved.

Then, based on the CCCP method [33], we replace the nonconvex part $-c_{f,b,t}^2$ with its first-order Taylor expansion to approximate constraint (18) at iteration $i+1$, which is given by

$$c_{f,b,t} - \left(c_{f,b,t}^{(i)2} + 2c_{f,b,t}^{(i)2}(c_{f,b,t} - c_{f,b,t}^{(i)})\right) \leq q_{f,b,t},$$
$$\forall f \in \mathcal{F}_t,\ b \in \mathcal{B}, \tag{20}$$

where $c_{f,b,t}^{(i)}$ denotes the value of $c_{f,b,t}$ obtained in the i-th iteration. Similarly, constraint (17a) at iteration $i+1$ is approximately expressed as

$$\frac{2\,Re\{(\mathbf{d}_{n,t}\mathbf{d}_{n,t}^H\mathbf{j}_{f,t}^{(i)})^H\mathbf{j}_{f,t}\}}{x_{n,t}^{(i)}} - \frac{|\mathbf{d}_{n,t}^H\mathbf{j}_{f,t}^{(i)}|^2}{(x_{n,t}^{(i)})^2}x_{n,t} \geq \gamma_{f,t},$$
$$\forall n \in \mathcal{G}_{f,t},\ f \in \mathcal{F}_t, \tag{21}$$

where $x_{n,t}^{(i)}$ and $\mathbf{j}_{f,t}^{(i)}$ denote the values of $x_{n,t}$ and $\mathbf{j}_{f,t}$ obtained in the i-th iteration, respectively. $Re\{x\}$ is the real part of complex value x. Consequently, problem $\mathcal{P}_S3$ can be tackled by iteratively solving the following convex problem, i.e.,

$$\mathcal{P}_S4: \max_{\substack{\mathbf{C}_t,\mathbf{j}_t,\{q_{f,b,t}\}\\ \{\gamma_{f,t}\},\{x_{n,t}\}}} \sum_{f\in\mathcal{F}_t} R'_{f,t} - \beta\sum_{b=0}^{B} R_{b,t}^{BH} - \lambda\sum_{b=0}^{B}\sum_{f\in\mathcal{F}_t} q_{f,b,t}$$

$$\text{s.t. } (11), (15), (16), (17\text{b}), (19), (20), \text{ and } (21),$$

which can be directly addressed by the interior point method (IPM) that adopts a standard optimization toolbox, such as CVX [34]. The process of solving problem $\mathcal{P}_S$ is elaborated in Algorithm 1, in which $\mathbf{A}'$ is the given cache placement matrix. According to [33], we should initialize penalty parameter λ with a small value λ^0 and increase it by a fixed rate μ in each iteration. In addition, λ_{max} is the maximum penalty value and should be sufficiently large. Since slack variables $\{q_{f,b,t}\}$ are not equal to zero exactly, a threshold ζ is applied to transfer $\mathbf{C}_t$ into binary. Specifically, when the value of $c_{f,b,t}$ obtained by solving $\mathcal{P}_S4$ is larger than ζ, we set $c_{f,b,t} = 1$, otherwise $c_{f,b,t} = 0$. Then, the other variables are further optimized according to the updated $c_{f,b,t}$.

Algorithm 1: The proposed algorithm for problem $\mathcal{P}_S$.

1: **Input:** Initialize $\mathbf{A}'$, $\mathbf{C}_t^{(0)}$, $\lambda^{(0)}$, λ_{max}, $\mu \geq 0$, t, and $\mathbf{j}_{f,t}^{(0)}$;
2: Set the iteration number $i \leftarrow 0$;
3: **repeat**
4: With the given cache placement matrix $\mathbf{A}'$, solve problem $\mathcal{P}_S4$ and denote the solution as $\{\mathbf{C}_t^*, \mathbf{j}_{f,t}^*, q_{f,b,t}^*, \gamma_{f,t}^*, x_{n,t}^*\}$;
5: Update $\mathbf{C}_t^{(i)} \leftarrow \mathbf{C}_t^*, \mathbf{j}_{f,t}^{(i)} \leftarrow \mathbf{j}_{f,t}^*$;
6: Update λ with $\min\{\mu\lambda, \lambda_{max}\}$;
7: Set $i \leftarrow i+1$;
8: **until** the convergence condition is satisfied;
9: **Output:** $\mathbf{C}_t^{(i)}$ and $\mathbf{j}_t^{(i)}$.

Recall that problems $\mathcal{P}_S2$ and $\mathcal{P}_S$ are equivalent. Meanwhile, after penalty parameter λ reaches the maximum value λ_{max}, the sequence of the results in iterations of solving $\mathcal{P}_S4$ will converge based on the theory of CCCP [33]. The simulation results in Section VII also demonstrate the convergence of Algorithm 1. Thus, Algorithm 1 can obtain a locally optimal solution for problem $\mathcal{P}_S$. It is worth noting that the computational complexity of solving problem $\mathcal{P}_S$ is dominated by the complexity of solving problem $\mathcal{P}_S4$. Specifically, we apply the basic elements of computational complexity analysis as in [35], in which the computational complexity of adapting IPM is $O(N_v^{3.5})$, where N_v is the number of variables. Therefore, the computational complexity of IPM for solving problem $\mathcal{P}_S$ is expressed as $O((F_t(2B+3+M_s+BM_b)+N)^{3.5}N_{i1})$, where $F_t(2B+3+M_s+BM_b)+N$ is the number of variables in problem $\mathcal{P}_S4$ and N_{i1} is the number of iterations to satisfy the convergence condition.

Remark 2: As a comparison, there are two drawbacks of adopting a semidefinite relaxation (SDR) and convex-concave procedure (SDR-CCP) based algorithm (considered in our previous work [1]) to solve the joint beamforming and clustering problem. For one thing, by adopting the SDR-CCP based algorithm, the number of variables increases to $F_t(1+(M_s+BM_b)^2)$, which is computationally expensive, especially when the numbers of ANs and transmit antennas are large. For another, the solution cannot guarantee the equivalence between the original problem and the recast problem. The result can be far from the optimal one owing to the relaxation of rank-one constraint.

VI. Proposed Algorithms for Long-Term Cache Placement Subproblem

In this section, we first propose a centralized algorithm to address the long-term cache placement problem, $\mathcal{P}_L$. To reduce the computational complexity, we then propose a distributed alternating algorithm based on the BCD to efficiently tackle problem $\mathcal{P}_L$.

A. Proposed Centralized Algorithm

To cope with nonconvex problem $\mathcal{P}_L$, we first reformulate it in a similar reformulation manner with problem $\mathcal{P}_S$ as follows:

$$\mathcal{P}_L1: \max_{\mathcal{E}_t,\mathbf{A}} \frac{1}{T}\sum_{t\in\mathcal{T}'}\left(\sum_{f\in\mathcal{F}_t} R'_{f,t} - \beta\sum_{b=0}^{B}\sum_{f\in\mathcal{F}_t} z_{f,b,t}\right)$$
$$\text{s.t. } (9a)-(9b),$$
$$(17a)-(17c), \quad \forall t\in\mathcal{T}',$$

where $\mathcal{E}_t=\{\{\mathbf{C}_t\},\{\mathbf{j}_t\},\{\gamma_{f,t}\},\{x_{n,t}\},\forall t\in\mathcal{T}'\}$ is the set of variables, $z_{f,b,t}\triangleq c_{f,b,t}(1-a_{f,b})m_f, \forall t\in\mathcal{T},'\ f\in\mathcal{F}_t,\ b\in\mathcal{B}$. Note that the objective function in problem $\mathcal{P}_L1$ is nonconvex due to the bilinear product term, $c_{f,b,t}(1-a_{f,b})$. To overcome the nonconvexity challenge, inspired by [36], we introduce a nonnegative auxiliary variable $\theta_{f,b,t}$ to substitute $c_{f,b,t}(1-a_{f,b})$ in the objective function. To tighten this substitution, the following linear constraints are added:

$$\theta_{f,b,t}\le c_{f,b,t},\ \theta_{f,b,t}\ge c_{f,b,t}-a_{f,b}, \tag{22}$$
$$\theta_{f,b,t}+a_{f,b}\le 1,\ \theta_{f,b,t}\ge 0. \tag{23}$$

It can be verified that $c_{f,b,t}(1-a_{f,b})$ can be equivalently replaced by $\theta_{f,b,t}$ under constraints (22) and (23). In this regard, the backhaul traffic generated by AN b in each transmission frame t can be rewritten as $R_{b,t}^{BH}=\sum_{f\in\mathcal{F}_t}\theta_{f,b,t}m_f$. Accordingly, problem $\mathcal{P}_L1$ can be equivalently represented as

$$\mathcal{P}_L2: \max_{\mathcal{E}_t,\mathbf{A},\{\theta_{f,b,t}\}} \frac{1}{T}\sum_{t\in\mathcal{T}'}\left(\sum_{f\in\mathcal{F}_t} R'_{f,t} - \beta\sum_{b=0}^{B} R_{b,t}^{BH}\right)$$
$$\text{s.t. } (9a)-(9b),$$
$$(17a)-(17c), \quad \forall t\in\mathcal{T},'$$
$$(22) \text{ and } (23), \quad \forall t\in\mathcal{T},'\ f\in\mathcal{F}_t,\ b\in\mathcal{B}.$$

Problem $\mathcal{P}_L2$ is still nonconvex in terms of the same nonconvex constraints (17a) and (14) as in problem $\mathcal{P}_S2$ during each transmission frame t. To this end, the P-CCCP based algorithm proposed in Section V can also be adopted to iteratively solve problem $\mathcal{P}_L2$. Thus, in the i-th iteration, we solve the following approximated problem:

$$\mathcal{P}_L3: \max_{\substack{\mathcal{E}_t,\mathbf{A},\\ \{\theta_{f,b,t}\}}} \frac{1}{T}\sum_{t\in\mathcal{T}'}\left(\sum_{f\in\mathcal{F}_t} R'_{f,t} - \beta\sum_{b=0}^{B} R_{b,t}^{BH} - \lambda\sum_{b=0}^{B}\sum_{f\in\mathcal{F}_t} q_{f,b,t}\right)$$
$$\text{s.t. } (11),(15),(16),(17b),(19),(20), \text{ and } (21),\ \forall t\in\mathcal{T}',$$
$$(22) \text{ and } (23), \quad \forall t\in\mathcal{T},'\ f\in\mathcal{F}_t,\ b\in\mathcal{B}.$$

Problem $\mathcal{P}_L3$ is convex, and we can address it by IPM with CVX. Similar to $\mathcal{P}_S4$, the sequence of the results in iterations of solving $\mathcal{P}_L3$ can be converged according to the principle of CCCP [33]. The computational complexity of solving problem $\mathcal{P}_L$ is dominated by solving problem $\mathcal{P}_L3$. Thus, the computational complexity of IPM for solving problem $\mathcal{P}_L$ with centralized algorithm can be expressed as $O(((F_t(3B+4+M_s+BM_b)+N)T+F(B+1))^{3.5}N_{i2})$, where $((F_t(3B+4+M_s+BM_b)+N)T+F(B+1)$ is the number of variables in problem $\mathcal{P}_L3$ and N_{i2} is the number of iterations to reach the stopping criteria.

B. Proposed Distributed Algorithm

In Subsection VI-A, we proposed a P-CCCP based algorithm to iteratively solve problem $\mathcal{P}_L$ in a centralized manner. However, this algorithm may suffer from relatively high computational complexity when T is extremely large. Therefore, in this subsection, we propose a distributed alternating algorithm based on the BCD to efficiently solve problem $\mathcal{P}_L$.

Recall that the objective function in problem $\mathcal{P}_L$ is composed of the cumulative objective function of T short-term problem $\mathcal{P}_S$. In addition, based on the proposed two-step scheme in Section IV, the multicast beamformers and AN clustering matrix in each transmission frame can be obtained independently given the cache placement. Furthermore, long-term variable $\mathbf{A}$ is not coupled with short-term variables $\{\mathbf{w}_t\}$, $\{\mathbf{v}_t\}$, and $\{\mathbf{C}_t\}$ in all constraints of $\mathcal{P}_L$. Based on the above observations, the alternating method [37] combined with BCD is utilized to tackle problem $\mathcal{P}_L$. Specifically, we first divide all the variables into two blocks: block $\mathbf{A}$ and block $\Omega=\{\mathbf{w}_t,\mathbf{v}_t,\mathbf{C}_t,\forall t\in\mathcal{T}'\}$. Then the BCD based approach is devised to successively update the two blocks. Specifically, with the fixed block, $\mathbf{A}=\mathbf{A}^{(i)}$, we update block Ω by solving a total number of T problems $\mathcal{P}_S$ in parallel. With the fixed block, $\Omega=\Omega^{(i)}$, block $\mathbf{A}$ is dominated by the backhaul traffic function in problem $\mathcal{P}_L$, i.e., $R_{b,t}^{BH}=\sum_{f\in\mathcal{F}_t}c_{f,b,t}(1-a_{f,b})m_f$. Subsequently, we can update block $\mathbf{A}$ with fixed block Ω by solving the following convex problem:

$$\mathcal{P}_L4: \max_{\mathbf{A}} \ -\beta\frac{1}{T}\sum_{t\in\mathcal{T}'}\sum_{b=0}^{B}\sum_{f\in\mathcal{F}_t} c_{f,b,t}^{(i)}\left(1-a_{f,b}\right)m_f$$
$$\text{s.t. } (9a)-(9b),$$

which can be addressed by IPM with CVX.

The corresponding distributed algorithm for solving problem $\mathcal{P}_L$ is elaborated in Algorithm 2. Due to the contained step of updating block Ω by solving nonconvex problem $\mathcal{P}_S$, a suboptimal solution is obtained [38]. The distribution manner is characterized in Line 4, in which the computation task can be executed parallelly to significantly reduce the computational complexity. Thus, the computational complexity for solving problem $\mathcal{P}_L$ with the distributed algorithm can be expressed as $O(((F_t(2B+3+M_s+BM_b)+N)^{3.5}N_{i1}+(F(B+1))^{3.5})N_{i3})$, where N_{i3} is the number of iterations to reach the stopping criteria in Line 7. Furthermore, the sequence of the objective value of $\mathcal{P}_L4$ obtained by Algorithm 2 is increasing and upper bounded. To this end, Algorithm 2 can converge based on the monotone convergence theorem [29].

Algorithm 2: The proposed distributed algorithm for problem $\mathcal{P}_L$.

1: **Input:** Initialize $\mathbf{A}^{(0)}$, $\mathbf{C}_t^{(0)}$, $\lambda^{(0)}$, λ_{max}, $\mu \geq 0$, t, and $\mathbf{j}_{f,t}^{(0)}$;
2: Set the iteration number $i \leftarrow 0$;
3: **repeat**
4: Given cache placement matrix $\mathbf{A} = \mathbf{A}^{(i)}$, solve problems $\mathcal{P}_S4$, $\forall t \in \mathcal{T}'$ in parallel with Algorithm 1. The obtained solution is denoted by $\Omega^{(i+1)}$;
5: With the obtained block $\Omega = \Omega^{(i+1)}$, solve problem $\mathcal{P}_L4$. The obtained solution is denoted by $\mathbf{A}^{(i+1)}$;
6: Set $i \leftarrow i + 1$;
7: **until** the convergence condition is satisfied;
8: **Output:** $\mathbf{A}$.

VII. Performance Evaluation

A. Simulation Setup

In this section, we evaluate the performance of our proposed schemes with extensive simulation results. Consider a cache-enabled ISTN, where one LEO satellite and $B = 7$ BSs cooperatively provide services for $N = 14$ single-antenna users. Note that in practice, the LEO satellite has a broad coverage and provides diverse services for different user groups. In the simulation, we consider a relatively small area with 7 BSs. Accordingly, only a portion of the communication and cache resources of the LEO satellite is allocated for users in the considered area. Each BS is located at the center of a hexagonal cell with a radius of 800 meters, and all users are uniformly and independently distributed within the coverage area except for an inner circle of 50 m around each BS. The numbers of antennas of each BS and the satellite are set to be $M_b = 2$ and $M_s = 4$, respectively. The ISTN system operates at 2 GHz with a bandwidth of 10 MHz. The LEO satellite has an altitude of 1000 km with a peak transmission power of 46 dBm, and the maximum transmission power of each BS is set to be 43 dBm. For terrestrial communication channels, the path-loss is modeled by $PL(\text{dB}) = 148.1 + 37.6 \log_{10}(d)$, where d denotes the distance between the BS and the user in kilometers. The log-normal shadowing parameter is set to be 8 dB and the antenna gain of each BS is set to be 10 dBi. In addition, the small-scale fading is modeled as the normalized Rayleigh fading. For the wireless channel between users and the satellite, the antenna gain of the satellite is set to be 25 dBi, and the channel model is given in [30]. The power spectral density of noise is set to be -174 dBm/Hz.

In the simulation, the total number of contents in the library is 100, and all contents have the same size, i.e., $m_f = 125$ MByte. For simplicity, the relative caching capacities of BSs and the satellite are set to be 10% of the total content sizes. In practice, different users can have different content preferences. Thus, 4 types of preferences are considered in the simulation. Each user randomly selects its content preference among the 4 types. For each content preference, the content popularity rank order is randomly generated, and the value of δ is randomly chosen within interval [1, 2] [20]. Moreover, trade-off parameter β is initialized to be 0.005. The total amount of transmission frames in the long-term cache placement period is $T = 100$. Important simulation parameters are listed in Table I.

TABLE I
Simulation Parameters

Parameter	Value
The number of users, N	14
The number of contents, F	100
The number of BSs, B	7
Satellite altitude	1000 km
Carrier center frequency/Subcarrier bandwidth	2 GHz/10 MHz
The transmission power of BS, P_B^{max}	43 dBm
The transmission power of satellite, P_S^{max}	46 dBm
The number of antennas of BS, M_b	2
The number of antennas of satellite, M_s	4
The antenna gain of BS/satellite	10 dBi/25 dBi
The power spectral density of noise	-174 dBm/Hz
The fraction cache capacity of AN, x_b	10%
The size of content, m_f	125 MByte
The trade-off parameter, β	0.005
The number of transmission frames, T	100
The number of user preference types	4
The initial penalty parameter, λ	1
The maximum penalty parameter, λ_{max}	100

For notational simplicity, the two-step centralized scheme and distributed scheme are abbreviated to TSC and TSD, respectively. To evaluate the effectiveness of the proposed schemes, two common heuristic caching strategies and the scheme proposed by [1] are adopted as the benchmarks. For fairness consideration, the corresponding short-term transmission designs for the heuristic caching strategies are based on Algorithm 1. To demonstrate the benefits of cooperative transmission in the ISTN, a non-cooperative transmission scheme is also considered as a benchmark. The four benchmarks are described as follows.

- *Uniform Caching (UC)*: Each BS and the satellite uniformly cache the same fraction x_b of all contents regardless of content popularity distribution and users' preferences.
- *Popularity-Aware Caching (PAC)*: Each BS and the satellite cache the most popular contents until reaching the cache storage capacity. In this scheme, the content popularity distribution and users' preferences are assumed to be known. Note that the cached contents in all BSs and the satellite are the same due to the same cache storage capacity.
- *Two-Step Non-Cooperative Transmission (TSNCT)*: In this scheme, each multicast group $\mathcal{G}_{f,t}$ can be served by only one AN (BS or the satellite). The corresponding cache placement strategy for TSNCT is based on the distributed algorithm.
- *Two-Step Distributed Scheme (TSD)* [1]: This is also a two-step scheme leveraging the historical information. Nevertheless, the short-term delivery subproblem is solved by the SDR-CCP based algorithm (considered in our previous work [1]).

B. Convergence Performance

Fig. 3 shows the convergence performance of the proposed P-CCCP based algorithm (Algorithm 1) with four different trials.

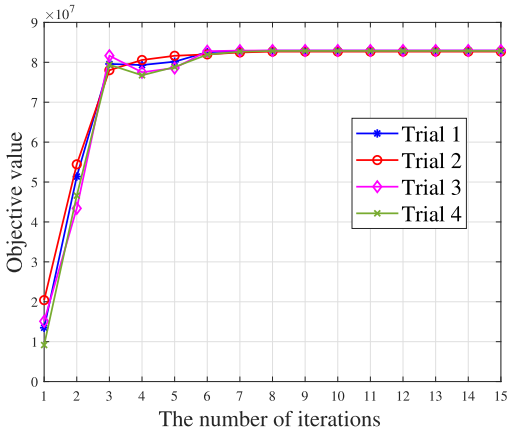


Fig. 3. Convergence performance of Algorithm 1.

We randomly initialize the elements of AN clustering matrix $\mathbf{C}_t^{(0)}$ within [0, 1]. Meanwhile, the penalty parameter,[5] λ, is initialized to be 1, λ_{max} is set to be 100, and threshold ζ is set to be 0.01. During each iteration, we increase λ by a fixed rate ($\mu = 4$) until it reaches the maximum value, λ_{max}. In general, the initialization may affect the results for many iterative algorithms. To verify the feasibility of Algorithm 1, multicast beamforming variables $\mathbf{w}_t$ and $\mathbf{v}_t$ are independently and randomly initialized within the feasible regions in different trials. As shown in Fig. 3, with different initialized multicast beamforming variables, the objective values exhibit an overall trend of increasing and quickly converge to the same value after 6 iterations. Furthermore, we observe that the objective values have small fluctuations between the third and fifth iterations. The reason is that the value of penalty parameter λ increases exponentially with the iterations and becomes sufficiently large at the 4-th iteration. Thus, the relaxed constraint, (18), is tightened rapidly, resulting in small fluctuations in the objective values.

C. Impact of Fractional Cache Capacity

In this subsection, we investigate the impact of fractional cache capacity x_b at ANs on the system performance in terms of the objective value of problem $\mathcal{P}$, network throughput, and total backhaul traffic. To be more practical, we choose small value for cache fraction because the cache storage of ANs is generally small compared with the total size of all contents in the library. As shown in Fig. 4(a), the proposed TSC scheme achieves a better performance compared with the proposed TSD scheme. However, the TSC scheme has a higher computational complexity as analyzed in Section VI. Therefore, when the network scale is small or the CP has powerful computing capability, the TSC scheme can be applied for better performance. In large-scale networks, the TSD scheme is more suitable and can achieve near-optimal performance with lower computational complexity.

[5]Note that the objective values of problem $\mathcal{P}_s$ and $\mathcal{P}_L$ are numerically large, which need a quite large penalty parameter. To make the problem more tractable, we divide the objective function by 10^7 in the simulation, and then restore it in the simulation results.

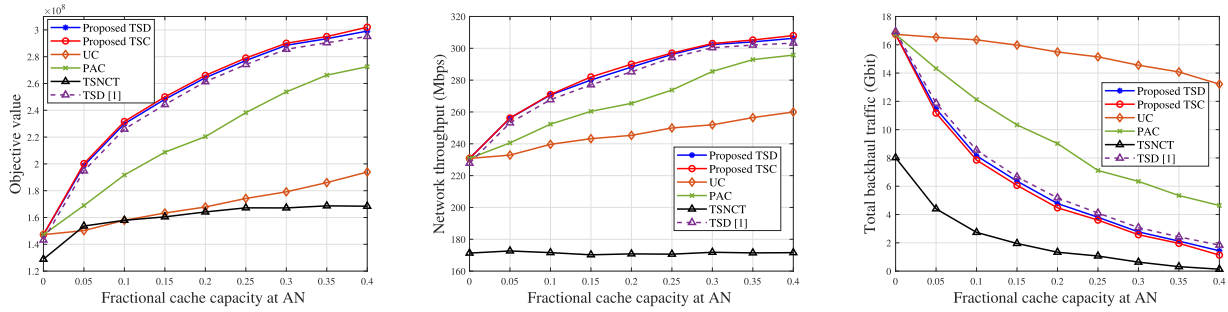
Moreover, the two proposed schemes are superior to UC, PAC, TSNCT, and TSD [1] schemes with 47.8%, 15.3%, 56.4%, and 2.5% on average increment in the objective value, respectively. As shown in Figs. 4(b) and (c), with a larger value of fractional cache capacity x_b, the network throughput increases whereas the performance of total backhaul traffic decreases. This is because caching more contents can provide more opportunities for cooperative transmission and further enhance the network throughput. Meanwhile, the increased fractional cache capacities significantly reduce the redundant transmissions in backhaul.

It is worth noting that the two proposed schemes averagely outperform the TSNCT scheme in network throughput by 65.4% since the proposed cooperative transmission scheme can efficiently enhance the received signal strength and decrease interference. Furthermore, the fractional cache capacity has little influence on the TSNCT scheme in network throughput, which is mainly due to the limited connectivity between users and ANs. However, the TSNCT scheme generates less amount of total backhaul traffic due to the non-cooperative transmission.

From the above simulation results, our proposed two-step schemes not only achieve much better performance in the network throughput, but also efficiently reduce the total backhaul traffic under different fractional cache capacities. The results demonstrate the benefits of our proposed joint optimization of cache placement, AN clustering, and multicast beamforming.

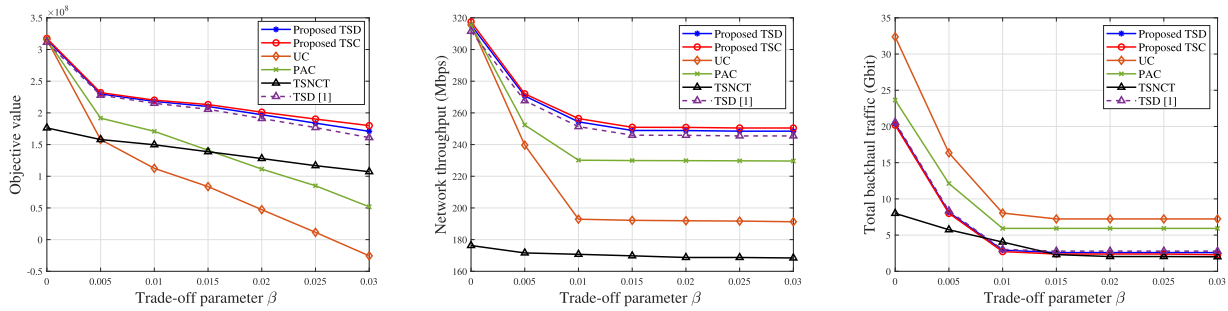
D. Impact of Trade-Off Parameter β

Fig. 5 shows the system performance in terms of the objective value of problem $\mathcal{P}$, network throughput, and total backhaul traffic under different values of trade-off parameter β. As shown in Fig. 5(a), the two proposed schemes have a close performance and outperform UC, PAC, TSNCT, and TSD [1] schemes with 121.1%, 45.6%, 59.5%, and 4.4% on average increment in the objective value, respectively. As shown in Fig. 5(b) and Fig. 5(c), when $\beta = 0$, all schemes have the maximum value of network throughput and total backhaul traffic. Meanwhile, UC, PAC, and the two proposed schemes achieve a similar network throughput in Fig. 5(b). This is because, the objective function is dominant by the network throughput which is determined by channel conditions and interference, regardless of the availability of contents in ANs. With the increase of β, both the network throughput and total backhaul traffic decrease since more emphasis is put on reducing the total backhaul traffic. It is noteworthy that, when β is larger than 0.015, the network throughput and total backhaul traffic achieved by all five schemes remain almost unchanged. This is intuitively plausible since the optimal AN clusters that provide services with the lowest backhaul traffic consumption have been identified. Furthermore, when β is larger than 0.01, the two proposed schemes significantly outperform the TSNCT scheme with an average network throughput augment of 47.9% with almost the same total backhaul traffic. The above results demonstrate that the proposed two-step schemes strike a superior balance between the network throughput and total backhaul traffic, thereby notably improving the system performance.



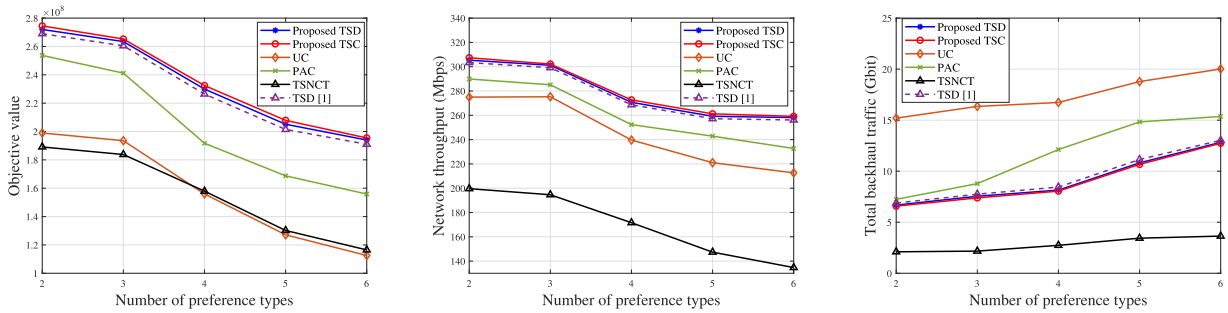
(a) Objective value (b) Network throughput (c) Total backhaul traffic

Fig. 4. The impact of fractional cache capacity at AN on the system performance. (a) Objective value. (b) Network throughput. (c) Total backhaul traffic.



(a) Objective value (b) Network throughput (c) Total backhaul traffic

Fig. 5. The impact of trade-off parameter β on the system performance. (a) Objective value. (b) Network throughput. (c) Total backhaul traffic.



(a) Objective value (b) Network throughput (c) Total backhaul traffic

Fig. 6. The impact of user preference types on the system performance. (a) Objective value. (b) Network throughput. (c) Total backhaul traffic.

E. Impact of User Preference

Fig. 6 shows the system performance in terms of the objective value of problem $\mathcal{P}$, network throughput, and total backhaul traffic with different numbers of user preference types. As shown in Fig. 6(a), the two proposed schemes exhibit comparable performance and outperform all the four benchmarks with a performance increment from 3.3% to 52.4%. In Fig. 6(b), the network throughput is in inverse proportion to the number of user preference types. This is because, when the number of user preference types increases, the requests from all users become more dispersive, and more multicast groups are generated, leading to higher inter-group interference and lower system performance. As shown in Fig. 6(c), with the increase of the number of user preference types, the total backhaul traffic becomes larger to satisfy the increment of groups.

The above simulation results demonstrate the superiority of the two proposed schemes under different numbers of user preference types. Although the user requests are highly dynamic and diversified, the two proposed schemes can learn from the historical users' request information to develop a wise transmission policy with optimized cache placement.

VIII. Conclusion

In this paper, we have investigated the joint design for long-term cache placement, short-term BS and satellite clustering, and multicast beamforming to improve the network utility in the ISTN. As the formulated optimization problem is anti-causal with mixed timescales, via leveraging historical information, we have proposed a two-step scheme to decompose the original problem into two subproblems which have been addressed by the

P-CCCP based algorithm and two different iterative algorithms, respectively. The proposed cache-enabled ISTN architecture can provide valuable insights into the cooperation between satellites and terrestrial BSs in the next-generation wireless networks, especially for content-centric applications. In the future, we will further explore the joint design considering the satellite mobility in the ISTN with LEO satellite constellations.

ACKNOWLEDGMENT

The authors would like to thank Jun Wang and Ruijin Sun for valuable comments and suggestions on the paper.

REFERENCES

[1] D. Han, H. Peng, H. Wu, W. Liao, and X. Shen, "Joint cache placement and content delivery in satellite-terrestrial integrated C-RANs," in *Proc. IEEE Int. Conf. Commun.*, 2021, pp. 1–6.
[2] H. Li, K. Ota, and M. Dong, "ECCN: Orchestration of edge-centric computing and content-centric networking in the 5G radio access network," *IEEE Wireless Commun.*, vol. 25, no. 3, pp. 88–93, Jun. 2018.
[3] "Mobile data traffic outlook," Accessed: Apr. 2021. [Online]. Available: https://www.ericsson.com/en/mobility-report/dataforecasts/mobile-traffic-forecast
[4] X. Shen *et al.*, "AI-assisted network-slicing based next-generation wireless networks," *IEEE Open J. Veh. Technol.*, vol. 1, no. 1, pp. 45–66, Jan. 2020.
[5] H. Wu, F. Lyu, C. Zhou, J. Chen, L. Wang, and X. Shen, "Optimal UAV caching and trajectory in aerial-assisted vehicular networks: A learning-based approach," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 12, pp. 2783–2797, Dec. 2020.
[6] C. Niephaus, M. Kretschmer, and G. Ghinea, "QoS provisioning in converged satellite and terrestrial networks: A survey of the state-of-the-art," *IEEE Commun. Surv. Tutor.*, vol. 18, no. 4, pp. 2415–2441, Oct./Dec. 2016.
[7] M. De Sanctis, E. Cianca, G. Araniti, I. Bisio, and R. Prasad, "Satellite communications supporting internet of remote things," *IEEE Internet Things J.*, vol. 3, no. 1, pp. 113–123, Feb. 2016.
[8] L. Zhen, A. K. Bashir, K. Yu, Y. D. Al-Otaibi, C. H. Foh, and P. Xiao, "Energy-efficient random access for LEO satellite-assisted 6G internet of remote things," *IEEE Internet Things J.*, vol. 8, no. 7, pp. 5114–5128, Apr. 2021.
[9] J. Li, K. Xue, D. S. L. Wei, J. Liu, and Y. Zhang, "Energy efficiency and traffic offloading optimization in integrated satellite/terrestrial radio access networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 4, pp. 2367–2381, Apr. 2020.
[10] X. Zhu, C. Jiang, L. Kuang, N. Ge, and J. Lu, "Non-orthogonal multiple access based integrated terrestrial-satellite networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2253–2267, Oct. 2017.
[11] M. Zhao, Y. Cai, M. Zhao, B. Champagne, and T. A. Tsiftsis, "Improving caching efficiency in content-aware C-RAN-based cooperative beamforming: A joint design approach," *IEEE Trans. Wireless Commun.*, vol. 19, no. 6, pp. 4125–4140, Jun. 2020.
[12] M. Tao, E. Chen, H. Zhou, and W. Yu, "Content-centric sparse multicast beamforming for cache-enabled cloud RAN," *IEEE Trans. Wireless. Commun.*, vol. 15, no. 9, pp. 6118–6131, Sep. 2016.
[13] E. Chen and M. Tao, "ADMM-based fast algorithm for multi-group multicast beamforming in large-scale wireless systems," *IEEE Trans. Commun.*, vol. 65, no. 6, pp. 2685–2698, Jun. 2017.
[14] X. Zhu, C. Jiang, L. Yin, L. Kuang, N. Ge, and J. Lu, "Cooperative multigroup multicast transmission in integrated terrestrial-satellite networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 5, pp. 981–992, May 2018.
[15] W. Wu *et al.*, "AI-native network slicing for 6G networks," *IEEE Wireless Commun.*, 2021, *arXiv:2105.08576*.
[16] Y. Zhang, L. Yin, C. Jiang, and Y. Qian, "Joint beamforming design and resource allocation for terrestrial-satellite cooperation system," *IEEE Trans. Commun.*, vol. 68, no. 2, pp. 778–791, Feb. 2020.
[17] T. X. Tran, D. V. Le, G. Yue, and D. Pompili, "Cooperative hierarchical caching and request scheduling in a cloud radio access network," *IEEE Trans. Mobile Comput.*, vol. 17, no. 12, pp. 2729–2743, Dec. 2018.
[18] J. Sung, M. Kim, K. Lim, and J. K. Rhee, "Efficient cache placement strategy in two-tier wireless content delivery network," *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1163–1174, Jun. 2016.
[19] Z. Qin, X. Gan, L. Fu, X. Di, J. Tian, and X. Wang, "Content delivery in cache-enabled wireless evolving social networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 10, pp. 6749–6761, Oct. 2018.
[20] X. Wu, Q. Li, X. Li, V. C. M. Leung, and P. C. Ching, "Joint long-term cache updating and short-term content delivery in cloud-based small cell networks," *IEEE Trans. Commun.*, vol. 68, no. 5, pp. 3173–3186, May 2020.
[21] Z. Yang, Y. Li, P. Yuan, and Q. Zhang, "TCSC: A novel file distribution strategy in integrated LEO satellite-terrestrial networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 5, pp. 5426–5441, May 2020.
[22] D. Jiang *et al.*, "QoE-aware efficient content distribution scheme for satellite-terrestrial networks," *IEEE Trans. Mobile Comput.*, to be published, doi: 10.1109/TMC.2021.3074917.
[23] B. Yang, Y. Wu, X. Chu, and G. Song, "Seamless handover in software-defined satellite networking," *IEEE Commun. Lett.*, vol. 20, no. 9, pp. 1768–1771, Sep. 2016.
[24] W. Wu *et al.*, "Dynamic RAN slicing for service-oriented vehicular networks via constrained learning," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 7, pp. 2076–2089, Jul. 2021.
[25] L. Li, D. Shi, R. Hou, R. Chen, B. Lin, and M. Pan, "Energy-efficient proactive caching for adaptive video streaming via data-driven optimization," *IEEE Internet Things J.*, vol. 7, no. 6, pp. 5549–5561, Jun. 2020.
[26] J. Choi and V. Chan, "Predicting and adapting satellite channels with weather-induced impairments," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 38, no. 3, pp. 779–790, Jul. 2002.
[27] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: Evidence and implications," in *Proc. IEEE Conf. Comput. Commun. 18th Annu. Joint Conf. IEEE Comput. Commun. Societies The Future is Now (Cat. No.99CH36320)*, 1999, pp. 126–134.
[28] J. Tang, T. Q. S. Quek, T. Chang, and B. Shim, "Systematic resource allocation in cloud RAN with caching as a service under two timescales," *IEEE Trans. Commun.*, vol. 67, no. 11, pp. 7755–7770, Nov. 2019.
[29] R. Sun *et al.*, "QoE-driven transmission-aware cache placement and cooperative beamforming design in cloud-RANs," *IEEE Trans. Veh. Technol.*, vol. 69, no. 1, pp. 636–650, Jan. 2020.
[30] A. Abdi, W. C. Lau, M. Alouini, and M. Kaveh, "A new simple model for land mobile satellite channels: First- and second-order statistics," *IEEE Trans. Wireless Commun.*, vol. 2, no. 3, pp. 519–528, May 2003.
[31] X. Xu, B. Du, and C. Wang, "On the bottleneck users for multiple-antenna physical-layer multicasting," *IEEE Trans. Veh. Technol.*, vol. 63, no. 6, pp. 2977–2982, Jul. 2014.
[32] K. G. Murty and S. N. Kabadi, "Some NP-complete problems in quadratic and nonlinear programming," *Math. Prog.*, vol. 39, no. 2, pp. 117–129, Jun. 1987.
[33] T. Lipp and S. Boyd, "Variations and extension of the convex-concave procedure," *Optim. Eng.*, vol. 17, no. 2, pp. 263–287, Jun. 2016.
[34] Y. Ye, M. Grant, and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.2" Accessed Apr. 2021. [Online]. Available: http://cvxr.com/cvx/
[35] Y. Ye, *Interior Point Algorithms: Theory and Analysis*, vol. 44. Hoboken, NJ, USA: Wiley, Inc., 2011.
[36] S. Lin, H. Ding, Y. Fang, and J. Shi, "Energy-efficient D2D communications with dynamic time-resource allocation," *IEEE Trans. Veh. Technol.*, vol. 68, no. 12, pp. 11 985–11 999, Dec. 2019.
[37] H. Peng, Q. Ye, and X. Shen, "Spectrum management for multi-access edge computing in autonomous vehicular networks," *IEEE Intell. Transp. Syst.*, vol. 21, no. 7, pp. 3001–3012, Jul. 2020.
[38] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex Optimization.* New York, NY, USA: Cambridge Univ. Press, 2004.

Dairu Han received the B.S. degree in 2015 from the Nanjing University of Science and Technology, Nanjing, China, where he is currently working toward the Ph.D. degree in aeronautical and astronautical science and technology. From November 2019 to May 2021, he was a Visiting Ph.D. Student with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. His research interests include satellite communication and cooperative transmission in integrated satellite-terrestrial networks.

Wenhe Liao received the Ph.D. degree from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 1996. He is currently a Professor with the Department of Aeronautical and Astronautical Engineering and has been the Vice President with the Nanjing University of Science and Technology, Nanjing, China, since 2010. His research has resulted in many papers, books, patents, systems, and equipment in the areas of satellite system design and high-end equipment design. He was the recipient of the Second Prize and the Third Prize of the National Science and Technology Progress Award of China. He is also an Executive Member of the Council of Chinese Institute of Astronautics and Chinese Institute of Aeronautics and Astronautics. He is a Member of Discipline Appraisal Group, Aeronautical and Astronautical Science and Technology, and Academic Degrees Committee of the State Council.

Haixia Peng (Member, IEEE) received the Ph.D. degrees in computer science and electrical and computer engineering from Northeastern University, Shenyang, China, in 2017, and the University of Waterloo, Waterloo, ON, Canada, in 2021, respectively. She is currently an Assistant Professor with the Department of Computer Engineering and Computer Science, California State University Long Beach, Long Beach, CA, U.S. Her current research interests include satellite-terrestrial vehicular networks, multi-access edge computing, resource management, and reinforcement learning. She was the TPC Member in IEEE VTC-fall 2016 & 2017, IEEE ICCEREC 2018, IEEE Globecom 2016-2021, and IEEE ICC 2017-2022 conferences.

Huaqing Wu (Member, IEEE) received the B.E. and M.E. degrees from the Beijing University of Posts and Telecommunications, Beijing, China, in 2014 and 2017, respectively, and the Ph.D. degree from the University of Waterloo, Waterloo, ON, Canada, in 2021. She is currently a Postdoctoral Research Fellow with the McMaster University, Burlington, ON, Canada. Her current research interests include vehicular networks with emphasis on edge caching, wireless resource management, space-air-ground integrated networks, and application of artificial intelligence (AI) for wireless networks. She was the recipient of the Best Paper Award at IEEE GLOBECOM 2018, Chinese Journal on Internet of Things 2020, and prestigious Natural Sciences and Engineering Research Council of Canada (NSERC) Postdoctoral Fellowship Award in 2021.

Wen Wu (Member, IEEE) recieved the B.E. degree in information engineering from the South China University of Technology, Guangzhou, China, in 2012, the M.E. degree in electrical engineering from the University of Science and Technology of China, Hefei, China, in 2015, and the Ph.D. degree in electrical and computer engineering from the University of Waterloo, Waterloo, ON, Canada, in 2019. He was a Postdoctoral Fellow with the Department of Electrical and Computer Engineering, University of Waterloo. He is currently an Associate Researcher with Frontier Research Center, Peng Cheng Laboratory, Shenzhen, China. His research interests include 6G networks, pervasive network intelligence, digital twin, and network virtualization.

Xuemin Shen (Fellow IEEE) received the Ph.D. degree in electrical engineering from Rutgers University, New Brunswick, NJ, USA, in 1990. He is currently a University Professor with the Department of Electrical and Computer Engineering, University of Waterloo, ON, Canada. His research interests include network resource management, wireless network security, Internet of Things, 5G and beyond, and vehicular networks. Dr. Shen is a registered Professional Engineer of Ontario, Canada, a Fellow of the Engineering Institute of Canada, a the Fellow of Canadian Academy of Engineering, a Fellow of the Royal Society of Canada, a Foreign Member of the Chinese Academy of Engineering, and a Distinguished Lecturer of the IEEE Vehicular Technology Society and Communications Society.

Dr. Shen was the recipient of the Canadian Award for Telecommunications Research from the Canadian Society of Information Theory (CSIT) in 2021, R.A. Fessenden Award in 2019 from IEEE, Canada, Award of Merit from the Federation of Chinese Canadian Professionals (Ontario) in 2019, James Evans Avant Garde Award in 2018 from the IEEE Vehicular Technology Society, Joseph LoCicero Award in 2015 and Education Award in 2017 from the IEEE Communications Society (ComSoc), Technical Recognition Award from Wireless Communications Technical Committee (2019) and AHSN Technical Committee (2013), Excellent Graduate Supervision Award in 2006 from the University of Waterloo, and the Premiers Research Excellence Award (PREA) in 2003 from the Province of Ontario, Canada. He was the Technical Program Committee Chair or Co-Chair of the IEEE GLOBECOM'16, IEEE INFOCOM'14, IEEE VTC'10 Fall, IEEE GLOBECOM'07, and the Chair of the IEEE ComSoc Technical Committee on Wireless Communications. Dr. Shen is the President Elect of the IEEE ComSoc. He was the Vice President of Technical & Educational Activities, Vice President for Publications, Member-at-Large on the Board of Governors, Chair of the Distinguished Lecturer Selection Committee, and a Member of the IEEE Fellow Selection Committee of the ComSoc. Dr. Shen was the Editor-in-Chief of the IEEE INTERNET OF THINGS JOURNAL, IEEE NETWORK, and *IET Communications*.